Embedded predictive analysis of misrepresentation risk in GLM ratemaking models

Michelle Xia, Lauren Anglin and Gary Vadnais

Northern Illinois University
intact INSURANCE

November 14, 2016
2016 CAS annual meeting

Funded by the Casualty Actuarial Society (CAS)

Motivation

- **Misrepresentation** (see, e.g., Winsor [1995]) is a type of insurance **fraud** when the applicant chooses to give a false statement on a risk factor that may affect the eligibility or rates of insurance (e.g., *traffic violation* history, annual *millage*, *use of vehicle*, *smoking* status and *age* in auto insurance).
- In practice, insurance companies usually do not verify information provided by the applicant.
- Due to the financial incentive, misrepresentation happens frequently.
- Misrepresentation is **unidirectional** and usually **unobserved**.

Ratemaking

In insurance ratemaking, actuaries determine auto insurance rates based on generalized linear models between **historical losses** and **risk factors** such as *use of vehicle*, *annual millage*, *traffic violation*, *claim history*, *age*, *location* and *smoking status*. For example, in personal auto ratemaking, we can specify a multiplicative model such as

$$\log(E(Y)) = use+millage+violation+claim+credit+age+gender+\cdots ,$$

where $E(Y)$ can be the **expected** collision loss for the individual in a policy year.

Misrepresentation and ratemaking

- In a traditional ratemaking model, misrepresentation will result in an **underestimation** of the risk/association. The estimated *relativity* will be smaller than that is indicated by the loss experience.
- Misrepresentation is usually **unobserved**, with the confirmed cases typically different to the unconfirmed ones (i.e., selection bias). Hence, from standard models we cannot estimate the *probability* of mispresentation or the correct *relativity* corresponding to the risk factor.
- When the risk factors are correlated, it could also lead to a **bias** in the estimation of the *relativity* for other risk factors.

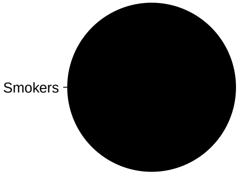
Misrepresentation mechanism

Suppose

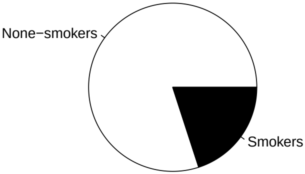
- There is a binary rating factor (e.g., smoking status) subject to misrepresentation
- $p =$ **probability of misrepresentation**
- $V =$ **true** binary risk status that we are not able to observe
- $V^* =$**observed** variable with a certain probability of misrepresentation
- We can write the conditional probabilities as

$$\begin{aligned} P(V^* = 0 \,|\, V = 0) &= 1 \\ P(V^* = 0 \,|\, V = 1) &= p. \end{aligned} \tag{1}$$

Misrepresentation on smoking status



(a) Report smoking



(b) Report nonsmoking

Figure: Here, we usually do **not observe** the true status, hence **cannot directly learn** the probability of misrepresentation.

Simplified example on smoking and health claim

- Suppose the smoking status (V) is the only risk factor that will affect the **severity** of a health insurance claim.
- We assume that the logarithm of loss (in thousands)

$$\begin{aligned} \log(Y) &\sim N(1,1) \quad \text{when V=0} \\ \log(Y) &\sim N(5,1) \quad \text{when V=1.} \end{aligned} \tag{2}$$

- Now let us do an audience survey regarding the smoking status and health claim severity.

Audience survey on smoking and health claim

In order to avoid having no smoker in the audience, we are just going to use a makeup status as follows.

1. Randomly pick a **true** smoking status $V = Yes$ or $V = No$, write it down without saying it.
2. If $V = No$, then simply set your **observed** $V^* = No$. Write write it down without saying it.
3. If $V = Yes$, then pick a number between 1 to 10. If the number is smaller than 4 ($p = 0.3$), then pick the **observed** $V^* = No$ (misrepresent). Otherwise, set $V^* = Yes$ (true status). Write down your *observed* status V^*, but DONOT say it.
4. Pick a number between 1 between 24 and write it down. Now depending on whether your **true** status is $V = Yes$ or $V = No$, find your corresponding loss from the distribution table.

Ratemaking data structure

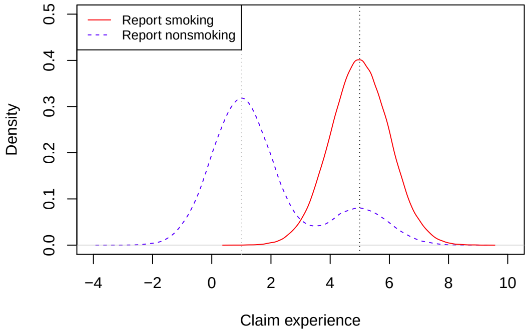


Figure: Loss experience by **reported** smoking status under **ratemaking** models, when comparing individuals with same **other risk characteristics**.

A general framework

Suppose $(Y \mid V, \mathbf{x})$ follows a distribution in the exponential family with a probability function $f_Y(y \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\varphi}, V, \mathbf{x})$ (e.g., in a regression model). Assume that the misrepresentation is **non-differential** (i.e., $(Y \perp V^* \mid V, \mathbf{x})$ and $(\mathbf{x} \perp V^* \mid V)$). In addition, assume $(\mathbf{x} \perp V)$, then we can write the conditional distribution of the observed variables as

$$\begin{aligned} f_Y(y \mid V^* = 1, \mathbf{x}) =& f_Y(y \mid \boldsymbol{\alpha}, \boldsymbol{\varphi}, V = 1, \mathbf{x}) \\ f_Y(y \mid V^* = 0, \mathbf{x}) =& q(\mathbf{x}) f_Y(y \mid \boldsymbol{\alpha}, \boldsymbol{\varphi}, V = 1, \mathbf{x}) \\ &+ (1 - q(\mathbf{x})) f_Y(y \mid \boldsymbol{\alpha}, \boldsymbol{\varphi}, V = 0, \mathbf{x}), \end{aligned} \tag{3}$$

where $q(\mathbf{x}) = \mathrm{P}(V = 1 | V^* = 0, \mathbf{x}) = \theta\, p(\mathbf{x}) / [1 - \theta(1 - p(\mathbf{x}))]$, $p(\mathbf{x}) = \mathrm{P}(V^* = 0 | V = 1, \mathbf{x})$ is the probability of misrepresentation, and θ is the binomial proportion for the true status V.

Health insurance model and assumptions

- For health insurance, we specify a **regression** structure that characterizes the relationship between **medical losses** and **true** risk profiles such as age, location and smoking status.
- We assume there is a latent mechanism on the misrepresentation of **smoking** status, and we know the **direction** of error.
- In addition, we can specify an **embedded predictive** model that associate the **probability** of misrepresentation to the **age** variable.

In more complicated cases, the **risk factors can be selected or tested**, like in the case of regular regression analysis.

Example: Claim frequency model

Denote V as the true status of prior condition, V^* as the **observed** smoking status with misrepresentation, $\mathbf{x}$ as a vector of K other correctly reported **risk factors**, and Y as the **number of health claims** in a policy year. Then we can use the negative binomial model given as

$$\begin{aligned}(Y \mid V, \mathbf{x}) &\sim negbin\,(\varphi, \beta_{V,\mathbf{x}}) \\ \log(\beta_{V,\mathbf{x}}) &= \alpha_0 + \alpha_1 V + \alpha_2 X_1 + \cdots + \alpha_{K+1} X_K \\ (V^* \mid V, \mathbf{x}) &\sim Bernoulli((1 - p(\mathbf{x}))V), \end{aligned} \tag{4}$$

where φ is the dispersion parameter, and $\beta_{V,\mathbf{x}}$ is the conditional mean of the negative binomial distribution given V and $\mathbf{x}$.

Here $f_Y(y \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\varphi}, V, \mathbf{x})$ is the *negative binomial pmf* with $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \cdots, \alpha_{K+1})$, $\boldsymbol{\beta} = \varnothing$, and $\boldsymbol{\varphi} = \varphi$.

Predictive analysis on misrepresentation

For the predictive analysis on the misrepresentation risk, we can embed a binary regression model in the models given in Equation (4). Denote $\mathbf{z}$ as a vector of rating factors that is a subset of $\mathbf{x}$ and $p(\mathbf{x}) = \mathrm{P}(V^* = 0 \mid V = 1, \mathbf{x})$, we can assume

$$\operatorname{logit}(p(\mathbf{x})) = \beta_0 + \mathbf{z}\boldsymbol{\beta}. \tag{5}$$

Using the Bayes's Theory, we can derive the the model for $q(\mathbf{x}) = P(V = 1|V^* = 0, \mathbf{x})$. That is,

$$\operatorname{logit}(q(\mathbf{x})) = \beta_0^* + \mathbf{z}\boldsymbol{\beta}, \tag{6}$$

where $\beta_0^* = \operatorname{logit}(\theta) + \beta_0$, β_0 is an intercept and the vector $\boldsymbol{\beta}$ contains the effects of the rating factors on the misrepresentation log odds in the logistic model on $p(\mathbf{x})$.

Three scenarios

We use the Poisson model as an example, and perform a simulation study for the three scenarios:

- Poisson model with an **additional** risk factor that is correctly measured
- Poisson model with **two** risk factors subject to misrepresentation
- Poisson model with an **embedded** model on the misclassification probability.

Three models compared

With a sample size of 1000, we compare the performance of three models:

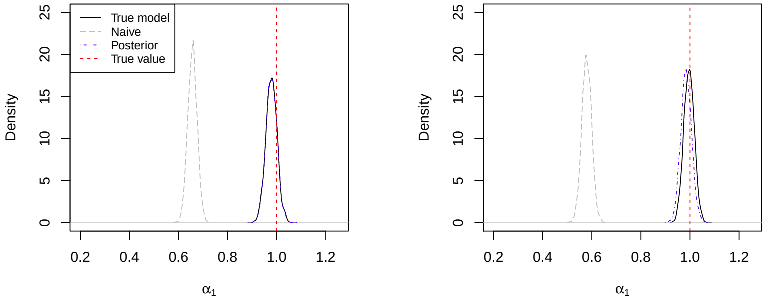
- **True** model where we assume the true status V is observed
- **Naive** model where we ignore the misrepresentation and use V^* in place of V
- **Posterior** model where we model the relationship of Y and V^* using the proposed method

Bayesian inference and non-informative priors

We use Bayesian inference based on Markov chain Monte Carlo (MCMC) simulations, and assume **non-informative** priors for all the parameters in the models.

$$
\begin{aligned}
\alpha_j &\sim N(0,\, 10) \\
p &\sim U(0,\, 1) \\
q &\sim U(0,\, 1) \\
\theta &\sim U(0,\, 1) \\
\beta_j &\sim N(0,\, 10).
\end{aligned}
$$

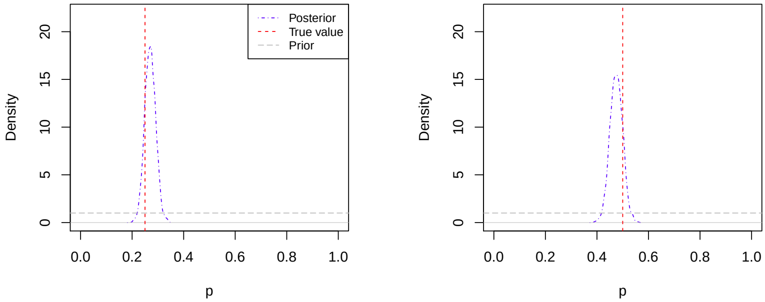
Additional risk factor: effect on misrepresented risk factor



(a) $p = 0.25$ (b) $p = 0.5$

Figure: Distribution of posterior samples for α_1 for the Poisson model.

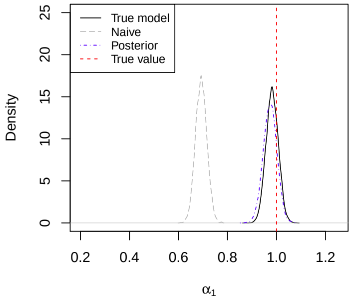
Additional risk factor: misrepresentation probability



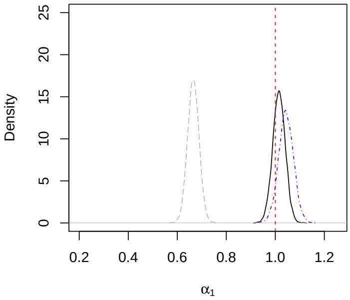
(a) $p = 0.25$ (b) $p = 0.5$

Figure: Distribution of posterior samples for p for the Poisson model.

Multiple risk factors: effect on misrepresented risk factor I



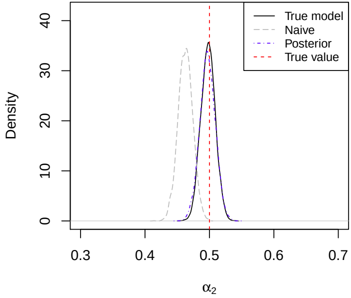
(a) $(p, q) = (0.25, 0.15)$



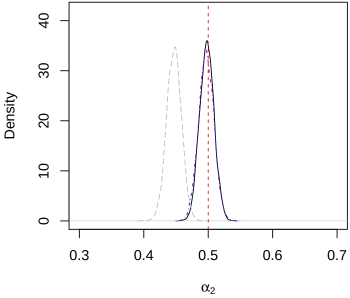
(b) $(p, q) = (0.35, 0.25)$

Figure: Distribution of posterior samples for α_1 for the Poisson model.

Predictive model: effect on correctly reported risk factor



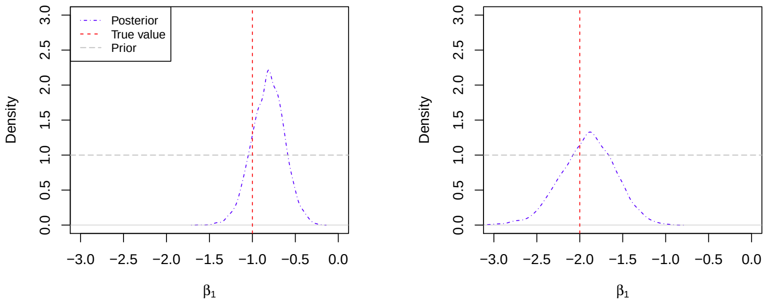
(a) $\beta_1 = -1$



(b) $\beta_1 = -2$

Figure: Distribution of posterior samples for α_2 for the Poisson model.

Predictive model: misreprentation model slope



(a) $\beta_1 = -1$ (b) $\beta_1 = -2$

Figure: Distribution of posterior samples for β_1 for the Poisson model.

Messages

- The naive model gives **biased** estimates on the **effect** α_1, with relativity being exp(effect).
- The proposed model gives results that are **similar** to those from the true model.
- The proposed model allows estimation of the **misrepresentation** probability, or the covariate **effects** on the misrepresentation probability when an **embedded** model is specified on the probability.

Medical Expenditure Panel Survey

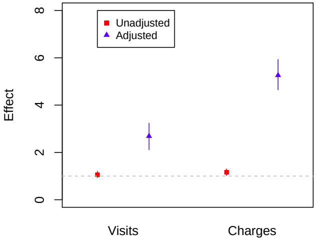
- The Medical Expenditure Panel Survey (MEPS) is a set of national surveys on the **frequency**, **cost** and source of **payment** for the health services that Americans use.
- For the case study, we include insured reference individuals **aged** from 18 to 60 inclusive, who are white and have a normal **BMI** between 18.5 to 30.
- The loss variables of interest Y are total **medical charges** (positive only) and number of **office-based visits**. The sample sizes for the two variables are 2948 and 3249, respectively.
- The variable V that is subject to misrepresentation is the **smoking** status.
- The additional covariate X is the **age** of the individual.
- In the **embedded** model, we assume that the probability of misrepresentation varies with **age**.

Objectives of study

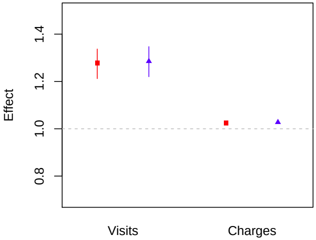
When modeling loss **frequency** (office-based visits, using *negative binomial* GLM) and **severity** (total medical charges, using *gamma* GLM),

- how does the adjustment of misrepresentation affect the *estimated relativity* for age and smoking status?
- how does the *probability* of misrepresentation in smoking status change with the age?
- given the age, what is the *probability* of misrepresentation for individuals who reported *nonsmoking*, i.e., $\mathrm{P}(V = 1 | V^* = 0)$?

Healthcare expense risk factors



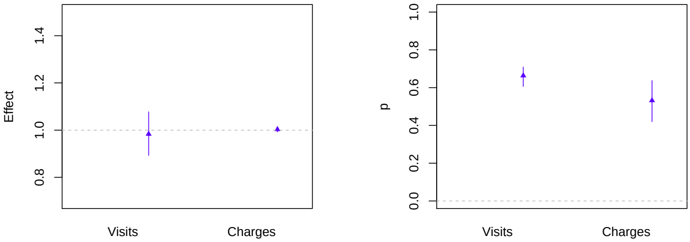
(a) Smoking



(b) Age

Figure: Credible intervals for the effect of **smoking** and **age**, for the office-based visits and total medical charges.

Misrepresentation risk factor



(a) Age (b) p

Figure: Credible intervals for **age** effect on **odds** of **misrepresentation**, and the **estimated misrepresentation probability** $p(x)$ for individuals at the **average age**.

Predictive model on misrepresentation probability

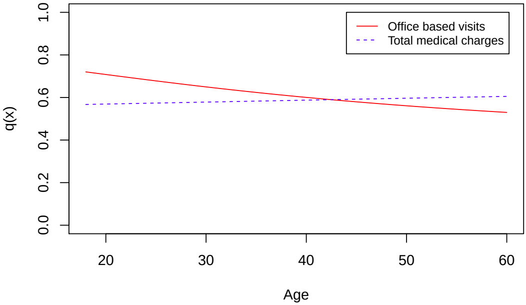


Figure: Predicted probability of misrepresentation for individuals who reported nonsmoking $q(x) = \mathrm{P}(V = 1 \mid V^* = 0, X = x)$.

How to use the model in GLM ratemaking?

In GLM ratemaking,

- the model uses **regular ratemaking data**, without requiring additional information on the misrepresentation.
- start with a **GLM ratemaking** model for loss frequency or severity, including various risk factors.
- embed a *latent* model on the **probability of misrepresentation**, with risk factors that may be predictive of the probability.
- based on the embedded model fitted on historical data, *predict* the **probability of misrepresentation** for each new policy where the applicant denies the risk status.

Thus, insurance companies may put more resources for investigating policies with a *higher probability* of misrepresentation, while ensuring the rates are *fair* with more accurate relativity estimated from the model.

Summary of work

- **Predictive analysis** on misrepresentation probability, e.g., by specifying a binomial **logistic** regression model on the misrepresentation probability p
- Inclusion of **additional** risk factors that are correctly measured
- Inclusion of **multiple factors** that are subject to misrepresentation

Take-home messages

- When unadjusted, misrepresentation in risk factors will result in an **underestimation** of the risk (e.g., relativity), in traditional GLM ratemaking models.
- Predictive analysis on the misrepresentation risk is possible by embedding a binomial **logistic** regression model on the probability of misrepresentation.
- The model can be implemented either using **Bayesian** analysis using MCMC, or **Maximum likelihood** estimation based on the Expectation Maximization algorithm.
- The method uses regular **ratemaking data**, without requiring additional information on the mirepresentation.
- The model provides more accurate rates, as well as predictive analysis on the misrepresentation probability.

Ongoing research

- Simulation study with other distributions
- Theoretical identification based on observable moments
- Misrepresentation on ordinal risk factors (Sun, et. al., 2016)
- Likelihood based inference with Expectation Maximization (EM) algorithm (Akakpo and Xia, 2016)

Acknowledgement

The work was supported by

- The 2016 Individual Grant by the Casualty Actuarial Society (CAS)
- The Student Engagement Fund (SEF) at Northern Illinois University (NIU)

Selected references

[1] Akakpo, R. and **Xia, M.** (2016). The Expectation Maximization algorithm for misrepresentation in insurance ratemaking models, *working paper*.

[2] Brockman, M.J. and Wright, T.S. (1992). Statistical motor rating: making effective use of your data. *Journal of the Institute of Actuaries*. 119: 457–543.

[3] Haberman, S. and Renshaw, A. E. (1996). Generalized linear models and actuarial science. *The Statistician*. 47: 407–436.

[4] **Xia, M.** and Gustafson, P. (2016). Bayesian regression models adjusting for unidirectional covariate misclassification. *The Canadian Journal of Statistics*, 44(2), 198–218.

[5] Sun, L. and **Xia, M.** (2016). Bayesian inference for unidirectional misclassification in ordinal covariates. *working paper*.

[6] Winsor, R. (1995). *Misrepresentation and non Disclosure on Applications for Insurance.* Blaney McMurtry LLP.

[7] **Xia, M.**, **Anglin, L.** and **Vadnais, G.** (2016). Embedded predictive analysis of misrepresentation risk in GLM ratemaking models. *working paper*.

Questions and comments

Thank You:)